Recitation 7
CS435: Introduction to Big Data

GTAs: Bibek R. Shrestha and Aaron Pereira
Email: cs435@cs.colostate.edu

October 12, 2018
Questions on PA2?
Introduction to PA3

Objectives:
- Estimate PageRank values under ideal conditions.
- Estimate PageRank values while considering dead-end articles (Taxation).
- Create Wikipedia Bomb
Datasets

- We have two datasets: Links Dataset and Titles Dataset.

- **Links-Simple-Sorted:**
  - Each line represents the outgoing links from a page.
  - The format of the lines is:
    - from1: to11 to12 to13
    - from2: to21 to22 to23
    - ...

- **Titles-Sorted:**
  - Each line represents title of Wikipedia article.

- To find the page title that corresponds to integer n, just look up the n-th line in the Titles-Sorted dataset.
Top 10 pages

- After the PageRank for the web graph converges (or after 25 iterations), show top 10 pages with highest page ranks.
- Then, join the updated ranks (top 10) with Titles dataset to get the corresponding titles.
Introduction to Apache Spark

- Download latest Apache Spark 2.3.2 binary (pre-build for Hadoop 2.7 and later): [Link]
- Go through its official documentation: [Link]
- Create Maven project and setup Spark on IDE to run WordCount program.
  - Refer Recitation 3: We did a demo on running Hadoop job on IDE.
- Try to setup Spark cluster on top of your Hadoop cluster. [Link]
In the next recitation

- Continue discussion on PA3